

Stochastic Models of Social Media Dynamics

Kristina Lerman, Aram Galstyan, Greg Ver Steeg
USC Information Sciences Institute
Marina del Rey, CA

Tad Hogg
Institute for Molecular Manufacturing
Palo Alto, CA

March 24, 2011

Abstract

A major challenge for designing future social media sites allowing a broader range of user actions is the difficulty of extrapolating from experience with current sites without first distinguishing correlations from underlying causal mechanisms leading to successful communities. The growing availability of data on user activities provides new opportunities to uncover correlations among user activity, contributed content and links among users. However, such correlations do not necessarily translate into methods for predicting outcomes or improving the productivity of the user communities that arise around social media. Instead, mechanistic models and intervention experiments provide a stronger basis for establishing causal mechanisms underlying the development of social media. In particular, stochastic models of large communities are well-suited to account for the large variation in user behavior, quality of contributed content, and effect of current events. Such models readily incorporate the structure of the web site, especially how content is presented to users, and thereby indicate the likely effects of design choices for new sites. We describe the ingredients of this approach, illustrate its use on Digg, a crowdsourced web site rating stories on current events [**Note:** mention any other examples], and its application to developing future social media.

[**Note:** Proposed outline for this position paper:

- introduction motivating use of mechanistic models for network science
- overview of stochastic modeling approach
- example: Digg (and perhaps others from your prior work? E.g. Flickr)

- discussion of extensions, future approaches (e.g., dynamics of community formation, integration between online and offline activities, including economics such as in online games; methods to promote trust and reputation,)

]

[**Note:** I've commented out the Digg figures – we can decide which, if any, are relevant as examples for this position paper; e.g., perhaps just a state diagram to illustrate the stochastic modeling approach and one of the model results such as predicted votes compared to observed votes for a few stories; and replace most or all of the equations and model parameters with a citation to the prior paper.]

1 Introduction

Social media are rapidly evolving with the creation of new services and growth of user communities. Currently, such sites generally provide a limited set of actions for their users: add and rate content and link to other users. To date, studies of existing social media sites mainly involve classifying the large available data sets according to features in the data, usually through a statistical regression based approach. Such methods can identify correlations among behaviors and suggest hypothesis for web site design choices leading to productive outcomes. However, such approaches are limited in their ability to identify causal mechanisms. Experiments, especially with multiple randomly-selected groups [42], are a more powerful approach. Specifically, Salganik et al. [42] experimentally measured the impact of content quality and social influence on the eventual popularity or success of cultural artifacts. They showed that while quality contributes only weakly to their eventual success, social influence,

or knowing about the choices of other people, is responsible for both the inequality and unpredictability of success. While revealing, such experiments are difficult to apply to existing large-scale user communities, especially where much of the behavior involves interaction among many users and external news events, such as Twitter or Digg.

Stochastic models provide another approach to identifying key mechanisms relating the design choices of social media web sites to the growth and performance of the user community. These models consider a few key features of the web site and users to define a set of states, and how users and web site content transition among these states probabilistically. Such mechanistic models of user behavior could aid development of future social media services by identifying key mechanisms leading to successful outcomes, particularly those that involve complicated feedback among user goals and effort, relevance and quality of contributed content, and relationships among users, especially those explicitly indicated on the site via its provision of links among users. These models become increasingly important as the complexity of social media web sites increase, allowing a wider diversity of user actions and feedbacks. In particular, as users gain a wider range of ways they can contribute and evaluate content, the community as a whole in effect becomes a computational platform where individual user actions contribute to identifying relevant content and improving it, e.g., as in the development of open source software and writing articles for Wikipedia. Utilizing such aggregated human computational abilities is likely to become ever more significant in the development of social media. As one example, with this approach we studied the social news aggregator Digg. We produced a model that helps explain — and predict [35] — the social voting patterns on Digg and related these aggregate behaviors to how Digg enables users to discover new content. One result from this modeling approach was the identification of key aspects of *homophily*, the commonality of users' interests indicated by links in the social network.

One application for stochastic models is to predicting popularity in social media. Popularity is not evenly distributed. Instead, a small number of users dominate the activity on the site and receive most of the attention of other users. The popularity of contributed items likewise shows extreme diversity. On the micro-blogging site

Twitter, for example, where a user's success may be defined in terms of the number of followers she has, a few users have upwards of a million followers, while the vast majority of users have a handful of followers. For example, relatively few of the four billion images on the social photo-sharing site Flickr are viewed thousands of times, while most of the rest are rarely viewed. Of the tens of thousands of new stories submitted daily to the social news portal Digg, only a handful become wildly popular, gathering thousands of votes, while most of the remaining stories never receive more than a single vote from the submitter herself. Among thousands of new blog posts every day, only a handful become widely read and commented upon. Success in social media is difficult to predict. Although early and late popularity, which can be measured in terms of user interest, e.g., votes or views, an item generates from its inception, are somewhat correlated [17, 45], we know little about what drives success. Does success derive mainly from an item's inherent quality [2], users' response to it [14], or some external factors, such as social influence [29, 31, 30]? Given the volume of new content, it is critically important to provide users with tools to help them sift through the vast stream of new content to identify interesting items in a timely manner, or least those items that will prove to be successful or popular. Accurate and timely prediction will also enable social media companies that host user-generated content to maximize revenue through differential pricing for access to content or ad placement, and encourage greater user loyalty by helping their users quickly find interesting new content. Moreover, models with predictive power based on underlying mechanisms can also suggest likely outcomes of design choices for new social media sites, especially those providing a wider range of user actions than typically available on current sites whose data will not include the proposed new features, preventing direct extrapolation from regression-based approaches.

The paper is organized as follows. Section 2 presents an overview of the stochastic modeling framework. In Section 3 we show how this modeling approach applies to the social news aggregator Digg. Section 6 generalizes this approach by describing aspects of future social media for which such modeling will likely be especially relevant.

2 Stochastic Models of Social Dynamics

Rather than account for the inherent variability of individuals, stochastic models focus on the macroscopic, or aggregate, behavior of the system, described by *average* quantities. In the context of social media, such quantities include average rate at which users contribute and rate content, and explicitly link to other users. Such macroscopic descriptions often have a simple form and are analytically tractable. Stochastic models do not reproduce the results of a single observation — rather, they describe the typical behavior.

These models are analogous to the approach used in statistical physics, demographics, epidemiology and macroeconomics where the focus is on relations among aggregate quantities, such as volume and pressure of a gas, population of a country and immigration, vaccination policy and fraction of a population contracting a disease or interest rates and employment.

We represent each individual entity, whether a user or a contributed content, as a stochastic process with a few states. This abstraction captures much of the individual complexity and environmental variability by casting individual’s actions as inducing probabilistic transitions between states. While this modeling framework applies to stochastic processes of varying complexity, for simplicity, we focus on processes that obey the Markov property, namely, a user whose future state depends only on her present state and the input she receives. A Markov process can be succinctly captured by a *state diagram* showing the possible states of the user and conditions for transition between those states. This approach is similar to compartmental models in biology [16]. For instance, in epidemiology such models track the progress of a disease as shifting individuals between states, or compartments, such as susceptible and infected.

We assume that all users have the same set of states, and that transitions between states depend only on the state and not the individual user. That is, the state captures the key relevant properties determining subsequent user actions. A choice of states to describe users results in grouping users in the same state into the same compartment for modeling. Then, the aggregate state of the system can be described simply by the *number* of individuals

in each state at a given time. That is, the system configuration at this time is defined by the occupation vector: $\vec{n} = (n_1, n_2, \dots)$ where n_k is the number of individuals in state k .

A key requirement for designing stochastic models is to ensure the state captures enough of the large variation in individual behavior to give a useful description of aggregate system properties. This is particularly challenging when individual activity follows a long-tail distribution, such as seen in some epidemics [38], as well as in social media web sites [8, 47]. In our case, including user link information as part of the state accounts for enough of this variation to provide reasonable accuracy, in particular significantly improving predictions compared to direct extrapolation of voting rates without accounting for the properties of the web site user interface.

The next step in developing the stochastic model is to summarize the variation within the collection of histories of changing occupation vectors with a probabilistic description. That is, we characterize the possible occupation vectors by the probability, $P(\vec{n}, t)$, the system is in configuration \vec{n} at time t . The evolution of $P(\vec{n}, t)$, governed by the Stochastic Master Equation [25], is almost always too complex to be analytically tractable. Fortunately we can simplify the problem by working with the average occupation number, whose evolution is given by the Rate Equation

$$\frac{d\langle n_k \rangle}{dt} = \sum_j w_{jk}(\langle \vec{n} \rangle) \langle n_j \rangle - \langle n_k \rangle \sum_j w_{kj}(\langle \vec{n} \rangle) \quad (1)$$

where $\langle n_k \rangle$ denotes the average number of users in state k at time t , i.e., $\sum_{\vec{n}} n_k P(\vec{n}, t)$ and $w_{jk}(\langle \vec{n} \rangle)$ is the transition rate from configuration j to configuration k when the occupation vector is $\langle \vec{n} \rangle$.

Using the average of the occupation vector in the transition rates is a common simplifying technique for stochastic models. A sufficient condition for the accuracy of this approximation is that variations around the average are relatively small. In many stochastic models of systems with large numbers of components, variations are indeed small due to many independent interactions among the components and the short tails of the distributions of these component behaviors. More elaborate versions of the stochastic approach give improved approximations when variations are not small, particularly due to cor-

related interactions [40] or large individual heterogeneity [39]. User behavior on the web, however, often involves distributions with long tails, whose typical behaviors differ significantly from the average [8, 47]. In this case we have no guarantee that the averaged approximation is adequate, even when aggregating the behavior of many users [43]. Instead we must test its accuracy for particular aggregate behaviors by comparing model predictions with observations of actual behavior, as we report below.

In the Rate Equation, occupation number n_k increases due to users' transitions from other states to state k , and decreases due to transitions from the state k to other states. The equations can be easily written down from the user state diagram. Each state corresponds to a dynamic variable in the mathematical model — the average number of users in that state — and it is coupled to other variables via transitions between states. Every transition must be accounted for by a term in the equation, with transition rates specified by the details of the interactions between users.

In summary, the stochastic modeling framework is quite general and requires only specifying the aggregate states of interest for describing the system and how individual user behaviors create transitions among these states. The modeling approach is best suited to cases where the users' decisions are mainly determined by a few characteristics of the user and the information they have about the system. These system states and transitions give the rate equations. Solutions to these equations then give estimates of how aggregate behavior varies in time and depends on the characteristics of the users involved.

3 Social News Portal Digg

[**Note:** depending on how much space we have after adding discussion of any other examples — e.g., Flickr — perhaps reduce the level of detail of this description of Digg, so we focus more on summarizing Digg as an example of the stochastic approach (especially giving an example of the rate equation to ground the fairly abstract discussion of rate equations in the general description of the stochastic method). This shorter summary of Digg results would keep the main focus of this position paper on application of stochastic modeling to future social media.]

With over 3 million registered users, the social news

aggregator Digg is one of the more popular news portals on the Web. Digg allows users to submit and rate news stories by voting on, or 'digging', them. There are many new submissions every minute, over 16,000 a day. Every day Digg picks about a hundred stories that it believes will be most interesting to the community and promotes them to the front page. Although the exact promotion mechanism is kept secret and changes occasionally, it appears to take into account the number of votes the story receives and how rapidly it receives them. Digg's success is fueled in large part by the emergent front page, which is created by the collective decision of its many users.

While the life cycle of each story may be drastically different from others, its basic elements are the same. These are specified by Digg's user interface, which defines how users post or discover new stories and interact with other users. A model of social dynamics has to take these elements into account when describing the evolution of story popularity.

3.1 User interface

A newly submitted story goes on the *upcoming* stories list, where it remains for a period of time, typically 24 hours, or until it is promoted to the front page, whichever comes first. The default view shows newly submitted stories as a chronologically ordered list, with the most recently submitted story at the top of the list, 15 stories to a page. To see older stories, a user must navigate to page 2, 3, *etc.* of the upcoming stories list. Promoted stories (Digg calls them 'popular') are also displayed as a chronologically ordered list on the *front pages*, 15 stories to a page, with the most recently promoted story at the top of the list. To see older promoted stories, user must navigate to page 2, 3, *etc.* of the front page. Figure 1 shows a screenshot of a Digg front page. Users vote for the stories they like by 'digging' them. The yellow badge to the left of each story shows its current popularity.

Digg allows users to designate friends and track their activities, i.e., see the stories friends recently submitted or voted for. The *friends interface* is available through the "Friends' Activity" link at the top of any Digg web page (see, for example, Fig. 1). The friend relationship is asymmetric. When user A lists user B as a *friend*, A can watch the activities of B but not vice versa. We call A the *fan* of B . A newly submitted story is visible in

Figure 1: Screenshot of the front page of the social news aggregator Digg.

the upcoming stories list, as well as to submitter’s fans through the friends interface. With each vote, a story becomes visible to the voter’s fans through the friends interface, which shows the newly submitted stories that user’s friends voted for.

Digg allows users to view the most popular stories from the previous day, week, month, or year. Digg also implements a social filtering feature which recommends stories, including upcoming stories, that were liked by users with a similar voting history. This interface, however, was not available at the time the data for our study was collected and hence is not part of the stochastic models described in this paper. Thus we examine a period of time when Digg had a relatively simple user interface, which simplifies the stochastic models.

3.2 Dynamics of popularity

By incorporating the various mechanisms through which web sites display content, stochastic models improve on predictions based on simply extrapolating from the early votes. Specifically, for one such site, the news aggregator Digg, we show how a stochastic model distinguishes the effect of the increased visibility due to the network from how interested users are in the content. We find a wide range of interest, distinguishing stories primarily of interest to users in the network (“niche interests”) from those of more general interest to the user community. This distinction is useful for predicting a story’s eventual popularity from users’ early reactions to the story.

By separating the impact of story quality and social influence on the popularity of stories on Digg, a stochastic model of social dynamics enables two novel applications: (1) estimating inherent story quality from the evolution of its observed popularity, and (2) predicting its eventual popularity based on users’ early reactions to the story. Specifically, to predict how popular a story will become, we use the early votes, even those cast before the story is promoted, to estimate how interesting it is to the user community. With this estimate, the model then determines, on average, the story’s subsequent evolution. We study these claims empirically on a sample of stories from

Digg. We show adjusting for the differing interests among voters based upon the social network improves predictions of popularity from early reactions of users.

We focus on modeling the behavior (i.e., votes received) of individual stories. Thus in our application of this approach, there is a different occupation vector for each story. For example, the states of a user with respect to a given story on Digg could be “has not seen the story,” “has seen the story but did not vote for it” and “has voted for the story.” The corresponding occupation vector has three elements, counting the number of users in each of these three compartments at a given time. As the story gains votes, users transition to the “has voted for the story” state, increasing the value of the corresponding element of the occupation vector. As described below, in our application of this approach to social media, we include the social network links of the users as part of the state and hence the occupation vectors we use have more than three elements.

While a story is in the upcoming stories list, it accrues votes slowly. If the story is promoted to the front page, it accumulates votes at a much faster pace. Figure 2(a) shows evolution of the number of votes for two stories submitted in June 2006. The point where the slope abruptly increases corresponds to promotion to the front page. The vast majority of stories are never promoted and, therefore, never experience the sharp rise in the number of votes that accompanies being featured on the front page. As the story ages, accumulation of new votes slows down [48], and after a few days the total number of votes received by a story saturates to some value. This value, which we call the final number of votes, gives a measure of the story’s success or *popularity*.

Popularity varies widely from story to story. Figure 2(b) shows the distribution of the final number of votes received by front page stories that were submitted over a period of about two days in June 2006. The distribution shows ‘inequality of popularity’: a handful of stories become very popular, accumulating thousands of votes, while most others only muster a few hundred votes. This distribution applies to front page stories only. Stories that are never promoted to the front page receive very

(a) (b)

Figure 2: Dynamics of social voting. (a) Evolution of the number of votes received by two front page stories in June 2006. (b) Distribution of popularity of 201 front page stories submitted in June 2006.

few votes, in many cases just a single vote from the submitter. In systems displaying such ‘long tailed’ distributions, extreme events, e.g., a story receiving many thousands of votes, occur much more frequently than would be expected if the underlying processes were Poisson or Gaussian in nature.

Long tails are ubiquitous features of human activity [4]. Examples include inequality of popularity of cultural artifacts, such as books and music albums [42], and in a variety of online behaviors [47], including tagging, where a few documents are tagged much more frequently than others, collaborative editing on wikis [28], and votes on a sample of more than 30,000 stories promoted to Digg’s front page over the course of a year [48].

While unpredictability of popularity is more difficult to verify than in the controlled experiments of Salganik et al., it is reasonable to assume that a similar set of stories submitted to Digg on another day will end with radically different numbers of votes. In other words, while the distribution of the final number of votes these stories receive will look similar to the distribution in Figure 2(b), the number of votes received by individual stories will be very different in the two realizations.

3.3 Data collection

We collected data for the study by scraping Digg’s Web pages in May and June 2006. The May data set consists of stories that were submitted to Digg May 25-27, 2006. We followed these stories by periodically scraping Digg to determine the number of votes stories received as a function of the time since their submission. We collected at least 4 such observations for each of 2152 stories, submitted by 1212 distinct users. Of these stories, 510, by 239 distinct users, were promoted to the front page. We followed the promoted stories over a period of several days, recording the number of votes the stories received. This May data set also records the location of the stories on the upcoming and front pages as a function of time.

The June data set consists of 201 stories promoted to

Figure 3: Voting rate (digs per hour) on front page stories at the end of June 2006. The indicated dates are the start of each day (0:00 GMT). The minimum in daily activity is around 9am GMT. Each point is the average voting rate for 100 successive votes.

the front page between June 27 and 30, 2006. For each story, we collected the names of its first 216 voters.

We focus on the early stages of story evolution – from submission until shortly after promotion – because the Digg social network has a much larger effect on upcoming than front page stories due to the much more rapid addition of stories to the upcoming list. This large influx of stories makes it difficult for users to find a new story before it becomes hidden by the arrival of more stories. In this case, enhanced visibility via the network for fans of the submitter or early voters is particularly important, and a model of social dynamics has to account for it. In light of these observations, and for speeding up data collection, we focus on the early votes for stories.

Activity on Digg varies considerably over the course of a day, as seen in Fig. 3. Adjusting times by the cumulative activity on the site accounts for this variation and improves predictions [45]. We define the “Digg time” between two events (e.g., votes on a story) as the total number of votes on front page stories during the time between those events. This behavior is similar to that seen in an extensive study of front page activity in 2007 [45], and as in that study we scale the measure by defining a “Digg hour” to be the average number of front page votes in an hour, which is 2500 for our data set.

In addition to voter activity, we extracted a snapshot of the social network of the top-ranked 1020 Digg users as of June 2006. This data contained the names of each user’s friends and fans. Since the original network did not contain information about all the voters in our data, we augmented it in February 2008 by extracting names of friends of about 15,000 additional users. Many of these users added friends between June 2006 and February 2008. Al-

though Digg does not provide the time a new link was created, it lists the links in reverse chronological order and gives the date the friend joined Digg. By eliminating friends who joined Digg after June 30, 2006, we were able to reconstruct the fan links for all voters in our data. This data allows us to identify, for each vote, whether the user was a fan of any prior voter on that story, in which case the story would have appeared in the friends interface for that user.

Votes by fans account for 6% of the votes in the June data set and about 3% of the front page votes.

The data sets used in this and previous works were collected before Digg’s API was introduced. Scraping Web pages to extract data had several issues. First, data had to be manually cleaned to ensure consistency. Second, since vote time stamps were not available on the Web page, we had to supplement June 2006 data by using the Digg API in October 2009 to obtain the time of each vote, the final number of votes the story received, and the time of promotion. In the intervening time, however, some of the users had deleted their accounts. Since we could not easily resolve the time of the vote of an inactive user, we had to delete these users from the voters list. We believe that the small fraction of data lost in this manner (less than 8% of the data) does not adversely affect the modeling study.

4 A Model of Social Dynamics of Digg

[**Note:** For this paper, just one model of Digg as an example of a stochastic model is sufficient. For simplicity, I kept our earlier model (from ICWSM09) instead of the newer model (with niche interests).]

Underlying a stochastic model of social dynamics is a behavioral model of an individual Web user. The behavioral model accounts for choices a Web site’s user interface allows users. Detailed data about human activity that can be collected from social media sites such as Digg allow us to parameterize the models and test them by comparing their predictions to the observed collective dynamics.

A prior study of social dynamics of Digg [20] used a simple behavioral model that viewed each Digg user as a stochastic Markov process, whose state diagram with re-

Figure 4: State diagram of user behavior for a single story. A user starts in the \emptyset state at the left, may find the story through one of the three interfaces and may then vote on it. At a given time, the story is located on a particular page of either the upcoming or front page lists, not both. This diagram shows votes for a story on either page p of the front pages or page q of the upcoming pages. Only fans of previous voters can see the story through the friends interface. Users in the friends, front or upcoming states may choose to leave Digg, thereby returning to the \emptyset state (with those transitions not shown in the figure). Users reaching the “vote” state remain there indefinitely and can not vote on the story again. Parameters next to the arrows characterize state transitions.

spect to a single story is shown in Fig. 4. According to this model, a user visiting Digg can choose to browse the *front* pages to see the recently promoted stories, *upcoming* stories pages for the recently submitted stories, or use the *friends* interface to see the stories her friends have recently submitted or voted for. She can select a story to read from one of these pages and, if she considers it interesting, *vote* for it. The user’s environment, the stories she is seeing, changes in time due to the actions of all the users.

We characterize the changing state of a story by three values: the number of votes, $N_{\text{vote}}(t)$, the story has received by time t after it was submitted to Digg, the list the story is in at time t (*upcoming* or *front* page) and its location within that list, which we denote by q and p for upcoming and front page lists respectively.

With Fig. 4 as a modeling blueprint, we relate the users’ choices to the changes in the state of a single story. In terms of the general rate equation (Eq. 1), the occupancy vector \vec{n} describing the aggregate user behavior at a given time has the following components: the number of users who see a story via one of the front pages, one of the upcoming pages, through the friends pages, and number of users who vote for a story, N_{vote} . Since we are interested in the number of users who reach the vote state, we do not need a separate equation for each state in Fig. 4: at a given time, a particular story has a unique location on the upcoming or front page lists. Thus, for simplicity, we can group the separate states for each list in Fig. 4, and

consider just the combined transition for a user to reach the page containing the story at the time she visits Digg. These combined transition rates depend on the location of the story in the list, i.e., the value of q or p for the story. With this grouping of user states, the rate equation for $N_{\text{vote}}(t)$ is:

$$\frac{dN_{\text{vote}}(t)}{dt} = r(\nu_f(t) + \nu_u(t) + \nu_{\text{friends}}(t)) \quad (2)$$

where r measures how interesting the story is, i.e., the probability a user seeing the story will vote on it, and ν_f , ν_u and ν_{friends} are the rates at which users find the story via one of the front or upcoming pages, and through the friends interface, respectively.

In this model, the transition rates appearing in the rate equation depend on the time t but not on the occupation vector. Nevertheless, the model could be generalized to include such a dependence if, for example, a user currently viewing an interesting story not only votes on it but explicitly encourages people they know to view the story as well.

4.1 Story Visibility

Before we can solve Eq. 2, we must model the rates at which users find the story through the various Digg interfaces. These rates depend on the story's location in the list. The parameters of these models depend on user behaviors that are not readily measurable. Instead, we estimate them using data collected from Digg, as described below.

Visibility by position in list A story's visibility on the front page or upcoming stories lists decreases as recently added stories push it further down the list. The stories are shown in groups: the first page of each list displays the 15 most recent stories, page 2 the next 15 stories, and so on.

We lack data on how many Digg visitors proceed to page 2, 3 and so on in each list. However, when presented with lists over multiple pages on a web site, successively smaller fractions of users visit later pages in the list. One model of users following links through a web site considers users estimating the value of continuing at the site, and leaving when that value becomes negative [22]. This model leads to an inverse Gaussian distribution of

the number of pages m a user visits before leaving the web site,

$$e^{-\frac{\lambda(m-\mu)^2}{2m\mu^2}} \sqrt{\frac{\lambda}{2\pi m^3}} \quad (3)$$

with mean μ and variance μ^3/λ . This distribution matches empirical observations in several web settings [22]. When the variance is small, for intermediate values of m this distribution approximately follows a power law, with the fraction of users leaving after viewing m pages decreasing as $m^{-3/2}$.

To model the visibility of a story on the m^{th} front or upcoming page, the relevant distribution is the fraction of users who visit *at least* m pages, i.e., the upper cumulative distribution of Eq. 3. For $m > 1$, this fraction is

$$f_{\text{page}}(m) = \frac{1}{2} \left(F_m(-\mu) - e^{2\lambda/\mu} F_m(\mu) \right) \quad (4)$$

where $F_m(x) = \text{erfc}(\alpha_m(m-1+x)/\mu)$, erfc is the complementary error function, and $\alpha_m = \sqrt{\lambda/(2(m-1))}$. For $m = 1$, $f_{\text{page}}(1) = 1$.

The visibility of stories decreases in two distinct ways when a new story arrives. First, a story moves down the list on its current page. Second, a story at the 15th position moves to the top of the next page. For simplicity, we model these processes as decreasing visibility, i.e., the value of $f_{\text{page}}(m)$, through m taking on fractional values within a page, i.e., $m = 1.5$ denotes the position of a story half way down the list on the first page. This model is likely to somewhat overestimate the loss of visibility for stories among the first few of the 15 items on a given page since the top several stories are visible without requiring the user to scroll down the page.

List position of a story Fig. 5(a) shows how the page number of a story on the two lists changes in time for three randomly chosen stories from our data set. The behavior is close to linear when averaging over the daily activity variation (shown in Fig. 3). For simplicity in this model, we ignore this variation and take a story's page number on the upcoming page q and the front page p at time t to be [20]

$$p(t) = k_f(t - T_{\text{promotion}}) + 1 \quad (5)$$

$$q(t) = k_u t + 1 \quad (6)$$

where $T_{\text{promotion}}$ is the time the story is promoted to the front page (or ∞ if the story is never promoted) and the slopes are given in Table 1. For a given story, $p(t)$ is only defined for times $t \geq T_{\text{promotion}}$ and $q(t)$ for $t < T_{\text{promotion}}$. Since each page holds 15 stories, these rates are $1/15^{\text{th}}$ the submission and promotion rates, respectively.

Front page and upcoming stories lists Digg prominently shows the stories on the front page. The upcoming stories list is less popular than the front page. We model this fact by assuming a fraction $c < 1$ of Digg visitors proceed to the upcoming stories pages.

We use a simple threshold to model how a story is promoted to the front page. Initially the story is visible on the upcoming stories pages. If and when the number of votes a story receives exceeds a promotion threshold h , the story moves to the front page. This threshold model approximates Digg’s promotion algorithm as of May 2006, since in our data set we did not see any front page stories with fewer than 44 votes, nor did we see any upcoming stories with more than 42 votes. We take $h = 40$ as an approximation to the promotion algorithm.

Friends interface The friends interface allows the user to see the stories her friends have (i) submitted, (ii) voted for, and (iii) commented on in the preceding 48 hours. Although users can take advantage of all these features, we only consider the first two. These uses of the friends interface are similar to the functionality offered by other social media sites: e.g., Flickr allows users to see the latest images his friends uploaded, as well as the images a friend liked.

The fans of the story’s submitter can find the story via the friends interface. As additional people vote on the story, their fans can also see the story. We model this with $s(t)$, the number of fans of voters on the story by time t who have not yet seen the story. Although the number of fans is highly variable, the average number of additional fans from an extra vote when the story has N_{vote} votes is approximately

$$\Delta s = aN_{\text{vote}}^{-b} \quad (7)$$

where $a = 51$ and $b = 0.62$, as illustrated in Fig. 5(b), showing the fit to the *increment* in average number of fans per vote over groups of 5 votes as given in the data. Thus early voters on a story tend to have more new fans (i.e.,

fans who are not also fans of earlier voters) than later voters.

The model can incorporate any distribution for the times fans visit Digg. We suppose these users visit Digg daily, and since they are likely to be geographically distributed across all time zones, the rate fans discover the story is distributed throughout the day. A simple model of this behavior takes fans arriving at the friends page independently at a rate ω . As fans read the story, the number of potential voters gets smaller, i.e., s decreases at a rate ωs , corresponding to the rate fans find the story through the friends interface, ν_{friends} . We neglect additional reduction in s from fans finding the story without using the friends interface.

Combining the growth in the number of available fans and its decrease as fans return to Digg gives

$$\frac{ds}{dt} = -\omega s + aN_{\text{vote}}^{-b} \frac{dN_{\text{vote}}}{dt} \quad (8)$$

with initial value $s(0)$ equal to the number of fans of the story’s submitter, S . This model of the friends interface treats the pool of fans uniformly. That is we assume no difference in behavior, on average, for fans of the story’s submitter vs. fans of other voters.

In summary, the rates in Eq. 2 are¹:

$$\begin{aligned} \nu_f &= \nu f_{\text{page}}(p(t)) \Theta(N_{\text{vote}}(t) - h) \\ \nu_u &= c \nu f_{\text{page}}(q(t)) \Theta(h - N_{\text{vote}}(t)) \Theta(24\text{hr} - t) \\ \nu_{\text{friends}} &= \omega s(t) \end{aligned}$$

where t is time since the story’s submission and ν is the rate users visit Digg. The first step function in ν_f and ν_u indicates that when a story has fewer votes than required for promotion, it is visible in the upcoming stories pages; and when $N_{\text{vote}}(t) > h$, the story is visible on the front page. The second step function in ν_u accounts for a story staying in the upcoming list for at most 24 hours. We solve Eq. 2 subject to initial condition $N_{\text{vote}}(0) = 1$, because a newly submitted story starts with a single vote, from the submitter.

4.2 Model Parameters

The solutions of Eq. 2 show how the number of votes received by a story changes in time. The solutions de-

¹ $\Theta(x)$ is a step function: 1 when $x \geq 0$ and 0 when $x < 0$.

(a) (b)

Figure 5: (a) Current page number on the upcoming and front pages vs. time for three different stories. Time is measured from when the story first appeared on each page, i.e., time it was submitted or promoted, for the upcoming and front page points, respectively. (b) Increase in the number of distinct users who can see the story through the friends interface with each group of five new votes for the first 46 users to vote on a story. The points are mean values for 195 stories, including those shown in (a), and the curve is based on Eq. 7. The error bars indicate the standard error of the estimated means.

parameter	value
rate general users come to Digg	$\nu = 600$ users/hr
fraction viewing upcoming pages	$c = 0.3$
rate a voters' fans come to Digg	$\omega = 0.12$ /hr
page view distribution	$\mu = 0.6, \lambda = 0.6$
fans per new vote	$a = 51, b = 0.62$
vote promotion threshold	$h = 40$
upcoming stories location	$k_u = 3.60$ pages/hr
front page location	$k_f = 0.18$ pages/hr
story specific parameters	
interestingness	r
number of submitter's fans	S

Table 1: Model parameters.

pend on the model parameters, of which only two parameters — the story's interestingness r and number of fans the submitter has S — change from one story to another. Therefore, we fix values of the remaining parameters as given in Table 1.

As described above, we estimate some of these parameters (such as the growth in list location, promotion threshold and fans per new vote) directly from the data. The remaining parameters are not directly given by our data set (e.g., how often users view the upcoming pages) and instead we estimate them based on the model predictions. The small number of stories in our data set, as well as the approximations made in the model, do not give strong constraints on these parameters. We selected one set of values giving a reasonable match to our observations. For example, the rate fans visit Digg and view stories via the friend's interface, given by ω in Table 1, has 90% of the fans of a new voter returning to Digg within the next 19 hours. As another example of interpreting these parameter

Figure 6: Evolution of the number of votes received by six stories compared with model solution.

S	r	final votes
5	0.51	2229
5	0.44	1921
40	0.32	1297
40	0.28	1039
160	0.19	740
100	0.13	458

Table 2: Parameters for the example stories, listed in decreasing order of total votes received by the story and hence corresponding to the curves in Fig. 6 from top to bottom.

values, for the page visit distribution the values of μ and λ in Table 1 correspond to about 1/6 of the users viewing more than just the first page. These parameters could in principle be measured independently from aggregate behavior with more detailed information on user behavior. Measuring these values for users of Digg, or other similar web sites, could improve the choice of model parameters.

4.3 Results

The model describes the behavior of all stories, whether or not they are promoted to the front page. To illustrate the model results, we consider stories promoted to the front page. Figure 6 shows the behavior of six such stories. For each story, S is the number of fans of the story's submitter, available from our data, and r is estimated to minimize the root-mean-square (RMS) difference between the observed

votes and the model predictions. Table 2 lists these values.

Overall there is qualitative agreement between the data and the model, indicating that the features of the Digg user interface we considered can explain the patterns of collective voting. Specifically, the model reproduces three generic behaviors of Digg stories: (1) slow initial growth in votes of upcoming stories; (2) more interesting stories (higher r) are promoted to the front page (inflection point in the curve) faster and receive more votes than less interesting stories; (3) however, as first described in [29], better connected users (high S) are more successful in getting their less interesting stories (lower r) promoted to the front page than poorly-connected users. These observations highlight a benefit of the stochastic approach: identifying simple models of user behavior that are sufficient to produce the aggregate properties of interest.

The only significant difference between the data and the model is visible in the lower two lines of Fig. 6. In the data, a story posted by the user with $S = 100$ fans is promoted before the story posted by the user with $S = 160$ fans, but saturates at smaller value of votes than the latter story. In the model, the story with larger r is promoted first and gets more votes.

Thus while the stochastic model is primarily intended to describe typical story behavior, we see it gives a reasonable match to the actual vote history of individual stories. Nevertheless, there are some cases where individual stories differ considerably from the model, particularly where an early voter happens to have an exceptionally large number of fans, thereby increasing the story’s visibility to other users far more than expected. This variation, a consequence of the long-tail distributions involved in social media, is considerably larger than seen, for example, in most statistical physics applications of stochastic models. The effect of such large variations is an important issue to address when using stochastic models to predict the behavior of individual stories in social media.

Fig. 7 shows parameters required for a story to reach the front page according to the model, and how that prediction compares to the stories in our data set. The model’s prediction of whether a story is promoted is correct for 95% of the stories in our data set. For promoted stories, the correlation between S and r is -0.13 , which is significantly different from zero (p -value less than 10^{-4} by a randomization test). Thus a story submitted by a poorly connected user (small S) tends to need high inter-

Figure 7: Story promotion as a function of S and r . The r values are shown on a logarithmic scale. The model predicts stories above the curve are promoted to the front page. The points show the S and r values for the stories in our data set: black and gray for stories promoted or not, respectively.

Figure 8: Distribution of interestingness (i.e., r values) for the promoted stories in our data set compared with the best fit lognormal distribution.

est (large r) to be promoted to the front page [29].

Figure 8 shows the estimated r values for the 510 promoted stories in our data set have a wide range of interestingness to users. That is, even after accounting for the variation in visibility of the stories, there remains a significant range in how well stories appeal to users. Specifically, Fig. 9 shows these r values fit well to a lognormal distribution

$$P_{\text{lognormal}}(\mu, \sigma; r) = \frac{1}{\sqrt{2\pi} r \sigma} \exp\left(-\frac{(\mu - \log(r))^2}{2\sigma^2}\right) \quad (9)$$

where parameters μ and σ are the mean and standard deviation of $\log(r)$. For the distribution of interestingness values, the maximum likelihood estimates of the mean and standard deviation of $\log(r)$ equal to -1.67 ± 0.04 and 0.47 ± 0.03 , respectively, with the ranges giving the 95% confidence intervals. A randomization test based on the Kolmogorov-Smirnov statistic and accounting for the fact that the distribution parameters are determined from the data [12] shows the r values are consistent with this distribution (p -value 0.35). While broad distributions occur in several web sites [47], our model allows factoring out the effect of visibility due to the user interface from the overall distribution of votes. Thus we can identify variation in users’ inclination to vote on a story they see.

The model described in this section gives a reasonable qualitative account of how user behavior leads to stories’ promotion to the front page and the eventual saturation in the number of votes they receive due to their decreasing visibility. In the section below we show how additional properties of the interface and user population can be

Figure 9: Quantile-quantile plot comparing observed distribution of r values with the lognormal distribution fit (thick curve). For comparison, the thin straight line from 0 to 1 corresponds to a perfect match between the data and the distribution.

added to the model for a more accurate analysis of the aggregate behavior. For example, submitter’s fans may find the story more interesting than the general Digg audience, corresponding to different r values for these groups of users. In addition, we modeled users coming to Digg independently with uniform rates ν and ω . In fact, the rates vary systematically over hours and days [45] as shown in Fig. 3, and individual users have a wide range in time between visits [46]. In our model, this variation gives time-dependent values for ν , describing the rate users come to Digg, and k_f and k_u , which relate to the rate new stories are posted and promoted.

The ability of the stochastic approach to incorporate details of user behaviors based on information available on the web site illustrates its value in providing insights into how aggregate behavior arises from the users, in contrast to models that evaluate regularities in the aggregate behaviors [48]. In particular, user models can help distinguish aggregate behaviors arising from intrinsic properties of the stories (e.g., their interestingness to the user population) from behavior due to the information the web sites provides, such as ratings of other users and how stories are placed in the site, i.e., visibility. Stochastic models have both explanatory and predictive power.

5 Related work

The Social Web provides massive quantities of data about the behavior of large groups of people. Researchers are using this data to study a variety of topics, including detecting [1, 37] and influencing [15, 26] trends in public opinion, and dynamics of information flow in groups [49, 36].

The stochastic modeling approach applies to any social media site by matching the state diagram to the information on users and content displayed by the site. For example, this approach models a political discussion web site where users propose and discuss topics of current politi-

cal interest [9]. This site differs from Digg in not having an equivalent of Digg’s front page, so topics change their visibility more gradually than on Digg. Moreover, the site provides more variety in the types of links users can form, which allows users to separate social contacts from those they do not know personally but whose political views they find significant. Thus the details of topic visibility differ from those of Digg. Nevertheless, the stochastic modeling approach applies and shows similar behaviors among users and the distribution of interestingness among topics [21]. Stochastic models also describe the behavior of posts and comments on blogs [18].

Beyond social media, the stochastic framework has been extensively used in the compartmental models of the spread of a disease within a population. These models assume the population is composed of susceptible and infected individuals (SIS models), or susceptible, infected and recovered (SIR models) individuals. In their simplest form, these models assume that every individual is in contact with every other individual [5, 19], although more realistic models take into account the connectivity of individuals [27] and strong fluctuations in the connectivity [39]. The stochastic modeling framework was also applied to study collective behavior of multi-robot systems [34, 33, 3]. This approach represents simple reactive robots by Markov processes [32].

Several researchers examined the role of social dynamics in explaining and predicting distribution of popularity of online content. Wilkinson [47] found broad distributions of popularity and user activity on many social media sites and showed that these distributions can arise from simple macroscopic dynamical rules. Wu & Huberman [48] constructed a phenomenological model of the dynamics of collective attention on Digg. Their model is parameterized by a single variable that characterizes the rate of decay of interest in a news article. Rather than characterize evolution of votes received by a single story, they show the model describes the distribution of final votes received by promoted stories. Our model offers an alternative explanation for the distribution of votes. Rather than novelty decay, we argue that the distribution can also be explained by the combination of a non-uniform variations in the stories’ inherent interest to users and effects of user interface, specifically decay in visibility as the story moves to subsequent front pages. Such a mechanism can also explain the distribution of popularity

of photos on Flickr, which would be difficult to characterize by novelty decay. Crane & Sornette [14] analyzed a large number of videos posted on YouTube and found that collective dynamics was linked to the inherent quality of videos. By looking at how the observed number of votes received by videos changed in time, they could separate high quality videos, whether they were selected by YouTube editors or spontaneously became popular, from junk videos. This study is similar in spirit to our own in exploiting the link between observed popularity and content quality. However, while this, and Wu & Huberman study, aggregated data from tens of thousands of individuals, our method focuses instead on the *microscopic* dynamics, modeling how individual behavior contributes to the observed popularity of content. In [35] we used the simple model of social dynamics, reviewed in this paper, to predict whether Digg stories will become popular.

Researchers found statistically significant correlation between early and late popularity of content on Slashdot [24], Digg and YouTube [45]. Specifically, similar to our study, Szabo & Huberman [45] predicted long-term popularity of stories on Digg. Through large-scale statistical study of stories promoted to the front page, they were able to predict stories' popularity after 30 days based on their popularity one hour after promotion. Unlike our work, their study did not specify a mechanism for evolution of popularity, and simply exploited the correlation between early and late story popularity to make the prediction. Our work also differs in that we predict popularity of stories shortly after submission, long before they are promoted.

Several researchers [30, 7, 13] found that early diffusion of information across an interlinked community is a useful predictor of how far it will spread across the network in general. Both [30] and [13] exploited the anti-correlation between these phenomena to predict final popularity. Specifically, the former work used anti-correlation between the number of early fan votes and stories' eventual popularity on Digg to predict whether stories submitted by well connected users will become popular. That work exploited social influence only to make the prediction, and the results were not applicable to stories submitted by poorly connected users which were not quickly discovered by highly connected users. In contrast, the approach described in this paper considers effects of social influence regardless of the connectedness of the submit-

ter, and also accounts for story quality in making a prediction about story popularity. More generally, in Digg the visibility of stories to non-fans increases significantly and abruptly upon promotion to the front page, leading to most votes coming from users who are not part of the submitter's social network. Thus Digg provides limited scope for evaluating graph-based methods. Other studies of graph-based methods for prediction [7, 13] thus focus on other settings where social connections have a larger role in the aggregate behavior.

6 Discussion

[**Note:** perhaps combine with Conclusion section; emphasizing stochastic models use for helping design future social media, with more complex actions; not just analyzing existing ones]

Applying stochastic models based on the information provided to users by the web site provides mechanistic models to studies of social network dynamics. These models complement regression-based statistical data analysis by suggesting underlying mechanisms. We expect such models will become increasingly important as social media web sites provide wider ranges of user actions, thereby making it more difficult to identify causal relationship through regression methods alone due to the large number of potentially relevant variables.

Stochastic models can treat dynamic networks, e.g., to model the arrival of new users and the decreasing activity of users who lose interest in a web site or stop following users to whom they previously linked. In this case, instead of treating the network as a static graph, the model would need to account for nodes and edges entering and leaving the graph [44]. One such application is to the design of social media for specialized settings, e.g., improving information flow and expertise sharing within organizations when highly active users leave the organization [10]. A key issue for such dynamic graphs is how rapidly the graph changes, with consequences for aggregate user behavior, compared to the rate at which nodes and links can be determined. For instance, in many web sites, users must register to participate and explicitly form links to other users. Thus the appearance of new nodes and links is well-defined. On the other hand, users who lose interest generally do not indicate that explicitly, e.g., by closing their account. Rather than become less active

or stop participating altogether, thereby requiring estimating when nodes or links leave the graph.

For the development of thriving user communities around social media, a key question is identifying user motivations for contributing, e.g., to Wikipedia [41]. This could aid in improving the appeal of web sites, especially to encourage productive participation by users with differing reasons for their interest and, conversely, reduce incentives for harmful actions, e.g., spamming other users by manipulating the apparent popularity of content. Thus part of the additional complexity of future social media sites will likely involve mechanisms for developing reputations on the quality of users' contributions to the community and the extent to which trust relationships can be transferred across links in the social graph of the community. Experiments would be particularly useful in this context, especially through the use of randomized trials [42]). Even better would be the ability to experimentally manipulate the motivations or utilities of users in controlled ways, not just the information they receive. Such manipulation is common in laboratory-based economic experiments [23] but is difficult in the context of the wide range of motivations for participating in social media. Sites whose design simplifies inference of utilities could facilitate including realistic utilities within stochastic models. Large online games are one such possibility [6, 7], especially those with an in-game economic component.

7 Conclusion

In the vast stream of new user-generated content, only a few items will prove to be popular, attracting a lion's share of attention, while the rest languish in obscurity. Predicting which items will become popular is exceedingly difficult, even for people with significant expertise. This prediction difficulty arises because popularity is weakly related to inherent content quality and social influence leads to an uneven distribution of popularity that is sensitive to the early choices of users in the social network. We described how stochastic models of user behavior on a social media web site can partially address this prediction challenge by quantitatively characterizing evolution of popularity. The model shows how popularity is affected by item quality and social influence. We evaluated the usefulness of this approach for the social news

aggregator Digg, which allows users to submit and vote on news stories. The number of votes a story accumulates on Digg shows its popularity. In earlier work we developed a model of social voting on Digg, which describes how the number of votes received by a story changes in time. In that model, knowing how interesting a story is to the user community, on average, and how connected the submitter is fully determines the evolution of the story's votes. This leads to an insight that a model can be used to predict story's popularity from the initial reaction of users to it. Specifically, we use observations of evolution of the number of votes received by a story shortly after submission to estimate how interesting it is, and then use the model to predict how many votes the story will get after a period of a few days. Model-based prediction outperforms other methods that exploit social influence only, and also correlation between early and late votes received by stories. We improved prediction by developing a more fine-grained model that differentiates between how interesting a story is to fans and to the general population.

These results demonstrate the applicability of the stochastic approach to social media, in spite of the large variations in user participation and interestingness of the content. A significant open question is the nature of the social influence on user behavior. In our model, the influence has two components: increased visibility of a story to fans due to the friends interface and the higher interestingness of the story to fans. This higher interestingness could be due to self-selection, whereby users become fans of people whose submissions or votes are of particular interest. Alternatively, users could be directly influenced by the activities of others [42], with the possibility that this influence depends not just on whether friends vote on a story but also how many friends do so [11]. Other challenging open questions including identifying common mechanisms underlying the observed regularities, accounting for time-dependent changes in the web site and user community, and extending the approach to a wider variety of web sites.

References

- [1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In

- Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.*
- [2] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 207–218, New York, NY, USA, 2008. ACM.
 - [3] W. Agassounon, A. Martinoli, and K. Easton. Macroscopic Modeling of Aggregation Experiments using Embodied Agents in Teams of Constant and Time-Varying Sizes. *Autonomous Robots*, 17(2-3):163–191, 2004.
 - [4] Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 2006.
 - [5] Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.
 - [6] William Sims Bainbridge. The scientific research potential of virtual worlds. *Science*, 317:472–476, 2007.
 - [7] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *EC '09: Proc. 10th ACM conference on Electronic commerce*, pages 325–334, 2009.
 - [8] Albert-Laszlo Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.
 - [9] Michael J. Brzozowski, Tad Hogg, and Gabor Szabo. Friends and foes: Ideological social networking. In *Proc. of the SIGCHI Conference on Human Factors in Computing (CHI2008)*, pages 817–820, NY, 2008. ACM Press.
 - [10] Michael J. Brzozowski, Thomas Sandholm, and Tad Hogg. Effects of feedback and peer pressure on contributions to enterprise social media. In *Proc. of the Intl. Conf. on Supporting Group Work (GROUP09)*, pages 61–70, NY, 2009. ACM Press.
 - [11] Damon Centola. The spread of behavior in an on-line social network experiment. *Science*, 329:1194–1197, 2010.
 - [12] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.
 - [13] Richard Colbaugh and Kristin Glass. Early warning analysis for social diffusion events. In *Proceedings of IEEE International Conferences on Intelligence and Security Informatics*, 2010.
 - [14] R. Crane and D. Sornette. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *Proc. of AAAI symposium on Social Information Processing*, Menlo Park, CA, 2008. AAAI.
 - [15] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. of KDD*, 2001.
 - [16] Stephen P. Ellner and John Guckenheimer. *Dynamic Models in Biology*. Princeton Univ. Press, Princeton, NJ, 2006.
 - [17] Vicente Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in Slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 645–654, New York, NY, USA, 2008. ACM.
 - [18] Michaela Gotz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. Modeling blog dynamics. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 26–33. AAAI, 2009.
 - [19] Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM REVIEW*, 42(4):599–653, 2000.
 - [20] Tad Hogg and Kristina Lerman. Stochastic models of user-contributory web sites. In *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 50–57, 2009.
 - [21] Tad Hogg and Gabor Szabo. Diversity of user activity and content quality in online communities. In

- Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 58–65. AAAI, 2009.
- [22] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280:95–97, 1998.
- [23] John Kagel and Alvin E. Roth, editors. *The Handbook of Experimental Economics*. Princeton Univ. Press, 1995.
- [24] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Proc. 5th Latin American Web Congress (LA-WEB 2007)*, 2007.
- [25] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science, Amsterdam, revised and enlarged edition, 1992.
- [26] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.
- [27] Jeffrey O. Kephart and Steve R. White. Directed-graph epidemiological models of computer viruses. *Security and Privacy, IEEE Symposium on*, 0:343, 1991.
- [28] A. Kittur, E. Chi, Bryan A. Pendleton, Bongwon Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of World Wide Web Conference*, 2006.
- [29] K. Lerman. Social networks and social information filtering on digg. In *Proc. of International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.
- [30] K. Lerman and A. Galstyan. Analysis of social voting patterns on digg. In *Proceedings of the 1st ACM SIGCOMM Workshop on Online Social Networks*, 2008.
- [31] K. Lerman and L. Jones. Social browsing on flickr. In *Proc. of International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.
- [32] K. Lerman, A. Martinoli, and A. Galstyan. A review of probabilistic macroscopic models for swarm robotic systems. In Sahin E. and Spears W., editors, *Swarm Robotics Workshop: State-of-the-art Survey*, number 3342 in LNCS, pages 143–152. Springer-Verlag, Berlin Heidelberg, 2005.
- [33] Kristina Lerman and Aram Galstyan. Mathematical model of foraging in a group of robots: Effect of interference. *Autonomous Robots*, 13(2):127–141, 2002.
- [34] Kristina Lerman, Aram Galstyan, A. Martinoli, and A. Ijspeert. A macroscopic analytical model of collaboration in distributed robotic systems. *Artificial Life Journal*, 7(4):375–393, 2001.
- [35] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of 19th International World Wide Web Conference (WWW)*, 2010.
- [36] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), 2007.
- [37] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Vanbriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, New York, NY, USA, 2007. ACM.
- [38] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438:355–359, 2005.
- [39] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 26(4):521–529, April 2002.

- [40] Manfred Opper and David Saad, editors. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA, 2001.
- [41] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proc. of the Intl. Conf. on Supporting Group Work (GROUP09)*, pages 51–60, NY, 2009. ACM Press.
- [42] M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854, 2006.
- [43] Didier Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Springer, Berlin, 2nd edition, 2004.
- [44] Christian Steglich, Tom A. B. Snijders, and Michael Pearson. Dynamics networks and behavior: Separating selection from influence. Technical report, Interuniversity Center for Social Science Theory and Methodology, July 2007.
- [45] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.
- [46] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E*, 73(3):036127+, March 2006.
- [47] Dennis M. Wilkinson. Strong regularities in online peer production. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, pages 302–309, New York, NY, USA, 2008. ACM.
- [48] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, November 2007.
- [49] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, June 2004.